Unsupervised M.L

Till now we have gone through all supervised learning techniques.

In case of Unsupervised ML techniques we have mainly ~~[illegible]~~ techniques:

- Clustrering techniqus
- KNN (K-Nearest neighbors)
- Anamaly detection
- PCA (Principle Component Analysis)
- Nural Network
- Independent Component Analysis
- Apriari Algorithm.

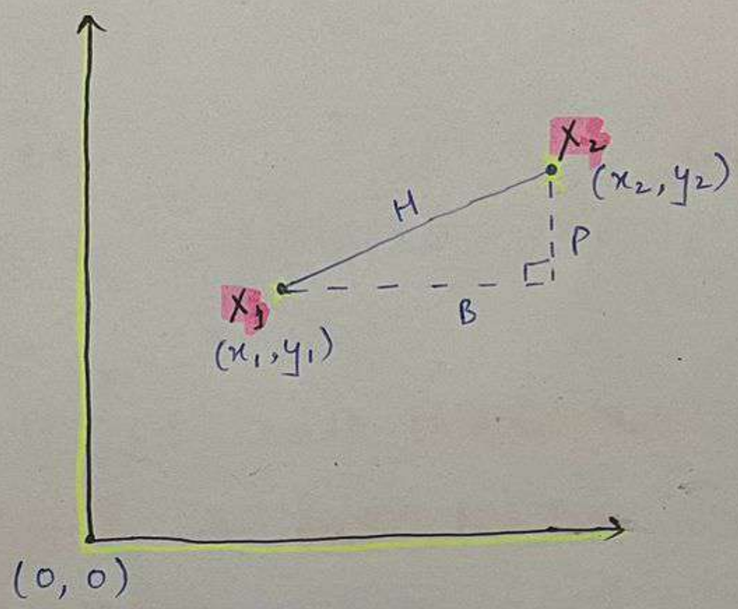
Clustering:

It's finding of structure or pattern in collection of Uncategarized data, then finding clusters (group) if it exist in the data. we can choose no.s of clusters we want to group our data.

Deciding datapoint for cluster to belong:

To decide which data point will belong to which cluster we use approach of Euclidian distance measure.

Euclidian distance measure:



distance b/w X_1 and X_2 can easily be calculated using pythagoras theorem:

$$\boxed{D(x_1, x_2) = H = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}$$

we also need to decide no.s of clusters bcz we just can't choose random no. as it will impact accuracy.

we use ELbow method to decide no.s of clusters we can choose.

let's consider below example to have bit idea of clustering and distance calculation.

S.no	Height	Weight	cluster	BWI
1	170	56	C_1	–
2	180	63	C_1	–
3	165	52		–
4	176	66		–
5	185	78	C_2	–
6	182	80		–

Initially we will choose any of point as centroid of our cluster and then will compare for which point will fall in which cluster.

let's we have consider two clusters, so we will have two centroid as data of S.no 2 and 5.

$C_1 \rightarrow$ S.no 2 , $C_2 \rightarrow$ S.no 5

Now using euclidian distance approach we will decide for rest points cluster either (C_1 or C_2) they will lie.

let's consider C_1 & C_2 as:

$C_1 \rightarrow (180, 63)$ | $C_2 \rightarrow (185, 78)$

let's consider any of the point from the table say point 1.

Then we will find distance of Point 1 (P_1) from both points C_1 & C_2

$$d(P_1, C_1) = \sqrt{(170-180)^2 + (56-63)^2}$$

$P_1 \rightarrow (170, 56)$ | $= 12.2$

$$d(P_1, C_2) = \sqrt{(170-185)^2 + (56-78)^2}$$

$$= 19.2$$

Now, distance $\boxed{d(P_1, C_1) < d(P_1, C_2)}$

Thus point P_1 will lies in the cluster of centroid C_1.

After it lies in the cluster of centroid C_1 we will then update the point.

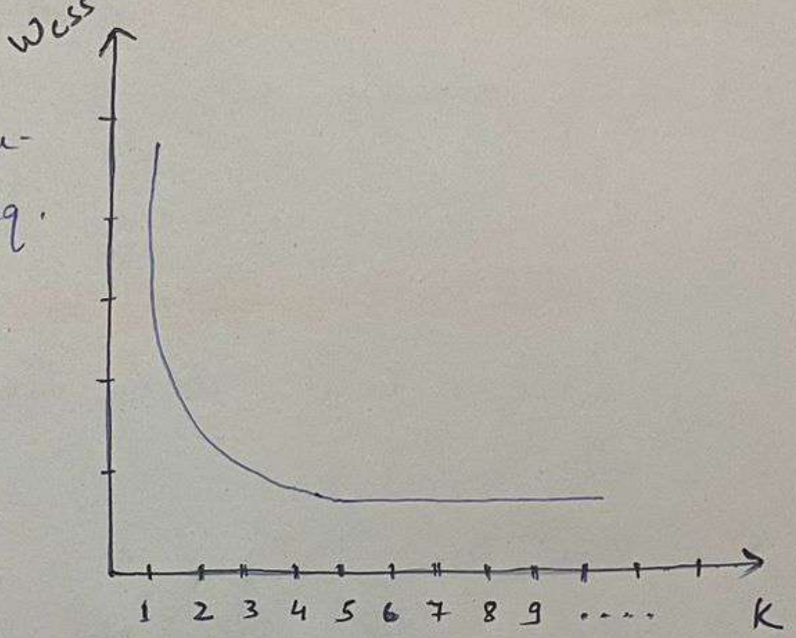
New centroind points of C_1 updated:

$$= \left(\frac{170+180}{2}, \frac{56+63}{2}\right)$$

$$= (175, 59.5)$$

ELBOW method :

we use Elbow method to decide no.s of clusters to have best prediction /model.



Wcss : With in cluster sum of sq.

k ; No.s of cluster.

$$\boxed{Wcss = \sum_{i=1}^{n} d(c_i, x_i)}$$

c_i : Centeroid position | x_i : considered datapoint position

Now, Elbow method states that till we have 5 no.s of clusters sum of Wcss is huge and after that either we increase no.s of cluster there is almost no diff in Wcss.

less Wcss will be, more accurate model will be.

we have two types of clusters of which we find Wcss

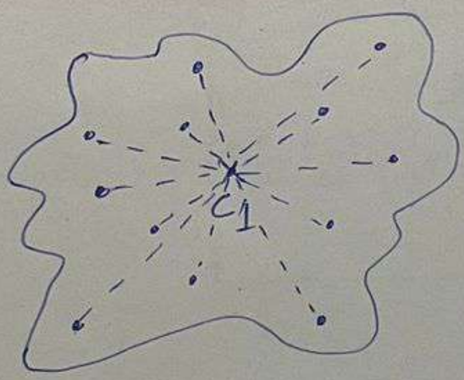
- Intra cluster.
- Inter cluster.

Intra cluster:

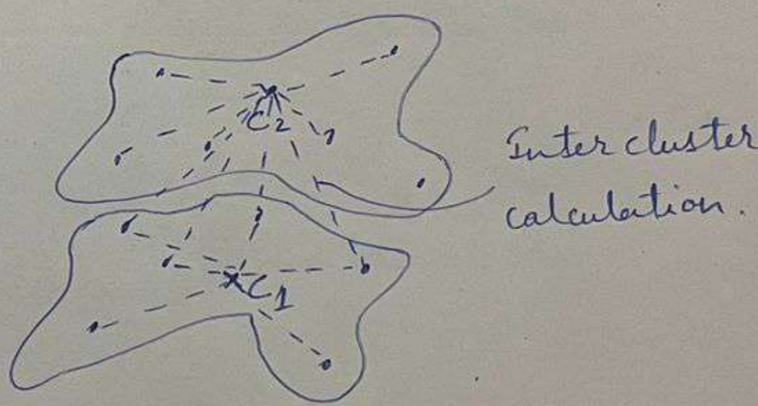
when all datapoints lies within the same cluster. we call it intra cluster.

Inter cluster:

when we have more than one cluster and datapoints lies in each cluster. It's Inter cluster.



1
Intra cluster



2
Inter cluster.

As we know $WCSS = \sum_{i=1}^{n} d(c_i, x_i)^2$

To find wcss square of sum is involved.
So more will be distance b/w centroid and datapoint much more will be sq. sum distance.

$$\boxed{WCSS_1 >> WCSS_2}$$

Thus increasing no.s of cluster decrease WCSS sum. But beyond $k = 5$ WCSS almost remains same.

Validation of no.s of clusters (k)

- once we made clusters we need to validate for it score/accuracy using following methods.
  - Dunn Index
  - Silhouettee score.

Dunn Index:

$$\boxed{\frac{\overbrace{\max(dis(x_i, x_j))}^{\text{Inter cluster}}}{\underbrace{\max(dis(y_i - y_j))}_{\text{Intra cluster.}}}}$$

Silhouettee score:

$$\boxed{\frac{b_i - a_j}{\max(\underbrace{b_i}_{\text{Inter cluster}} - \underbrace{a_j}_{\text{Intra cluster.}})}}$$

Silhouettee score lies $\boxed{\text{b/w } -1 \text{ to } 1}$